面向“为学习而测评”的纵向认知诊断模型[1]

詹沛达　潘艳方　李菲茗

(浙江师范大学教师教育学院)

摘　要　基于“为学习而测评”的理念，以促进学生学习为目的，客观量化学习现状并提供诊断反馈的测评模式日益受到重视。相比于横断认识诊断测评，纵向认知诊断测评更有利于实现促进学生发展的目标。为使国内学者系统性地了解纵向认知诊断模型，首先，依据建模逻辑将已有纵向认知诊断模型划分为基于潜在转换分析的和基于高阶潜在结构模型的两类；然后，逐一介绍和说明两类模型的理论基础和应用情景；最后，提炼出四个可进一步研究的议题。

关键词　认知诊断; 追踪研究; 潜在转换分析; 潜在类别分析; 纵向认知诊断模型

1. 引言

《教育部关于推进中小学教育质量综合评价改革的意见》(简称为《评价改革的意见》)指出当前中小学教育质量评价仍然“在评价内容上重考试分数忽视学生综合素质和个性发展，在评价方式上重最终结果忽视学校进步和努力程度，在评价结果使用上重甄别证明忽视诊断和改进。这些问题严重影响了学生的全面发展、健康成长，制约了学生社会责任感、创新精神和实践能力的培养。”教育的主要目的之一是促进学生的发展。良好的教育应遵循学生成长规律，应能够针对有个性差异的学生进行因材施教，发展每一个学生的优势潜能，进而促进学生主动获取和理解未知的知识内容。而“因材施教”的基本前提是要对学生现有的学习状况(e.g., 知识掌握情况和学习动机)及其在不同时间点上的发展(变化)情况(e.g., 知识掌握程度是否增加？学习动机是否增强？)有一个相对客观和准确的了解。因此，如何实现对学生现有的学习状况及其发展趋势进行客观且准确地测评是一个需要被关注的重要议题。

“有意义接受学习”理论(Ausubel, Novak, & Hanesian, 1968)强调已掌握知识对学习的重要性，认为有意义学习是学生将新知识纳入已有知识结构的过程。因此，客观量化学生的学习现状对促进学习具有重要作用。近些年，基于“为学习而测评(assessment for learning)”理念(Wiliam, 2011)，以促进学生学习为目的，客观量化学习现状并提供诊断反馈报告的测评模式日益受到重视。在众多测评方法中，

[1] 通讯作者：詹沛达, Email: pdzhan@gmail.com

认知诊断测评(cognitive diagnostic assessment, CDA)在近些年里得到了国内外学者的广泛关注。CDA 是指在心理与教育测量学中对个体认知过程、加工技能或知识结构(统称为属性)的诊断测评(Yang & Embretson, 2007)。作为一种将终结性评价和形成性评价相结合的综合评价形式(詹沛达, 陈平, 边玉芳, 2016)，CDA 的初衷是通过测评学生对属性的掌握状态为教师或干预者提供诊断反馈报告，进而帮助他们实施补救教学或有针对性的干预(Zhan, Jiao, & Liao, 2018)，其最终目的是为了更有效地促进学生的发展。因此，CDA 符合当前我国一些教育政策导向，比如，《评价改革的意见》中“将形成性评价与终结性评价相结合，注重考查学生进步的程度和学校的努力程度，改变单纯强调结果不关注发展变化的做法”；《基础教育课程改革纲要(试行)》中“改变课程评价过分强调甄别与选拔的功能，发挥评价促进学生发展、教师提高和改进教学实践的功能”的具体目标。总之，CDA 不仅有助于客观且准确地测评学生当前的认知过程和知识结构，还能进一步提供诊断反馈报告和补救教学建议，为促进学生发展奠定了理论基础。

“为学习而测评”理念强调反馈对促进学习的重要性，为判断 CDA 中诊断反馈与补救教学的成效，需依赖于沿学生成长轨迹收集的跨时间点的测评数据(i.e., 纵向数据或追踪数据)。作为 CDA 的核心技术环节，认知诊断模型(cognitive diagnosis model, CDM)与测验情景的匹配性或与数据的拟合性直接决定这测评结果的准确性和有效性。由于当前绝大多数 CDA 并未关注学生的发展问题而采用了横断测验设计，所以对于 CDM 的研究与应用也主要集中在对横断数据的分析。尽管已有个别研究尝试使用前后测设计[2]来测评学生的发展情况(e.g., Wu, 2018)，但仍使用了针对横断数据的传统 CDM (简称为“横断 CDM”)，即前测与后测分别使用相同的横断 CDM 进行数据分析。重复使用横断 CDM 分析纵向 CDA 数据的主要缺点是没有随时间的推移对模型参数进行同时校准，无法保证参数估计值在一个量尺上。鉴于横断 CDA 无法完全实现 CDA 的初衷及最终目标，纵向 CDA 逐渐受到研究者和实践者的关注。同时，如何更合理地分析纵向 CDA 数据(i.e., 开发纵向 CDM)已成为当前心理计量学的前沿议题。

近两年，在国际上，研究者们已经提出了一些不同的纵向 CDM (e.g., Chen, Culpepper, Wang, & Douglas, 2018; Hansen, 2013; Huang, 2017; Kaya & Leita, 2017; Li, Cohen, Bottge, & Templin, 2016; Studer, 2012; Wang, Yang, Culpepper, & Douglas, 2018; Wang, Zhang, Douglas, & Culpepper, 2018; Zhan, Jiao, Liao, & Li, 2019; Zhang & Wang, 2018)；而国内关于纵向 CDM 的研究尚为空白，仅有一篇文章简单介绍了 Li et al. (2016)及 Kaya 和 Leita (2017)的研究(张颖, 边玉芳, 2017)，尚不足以使国内学者对纵向 CDM 的

[2] 前后测设计等同于仅包含两个时间点的纵向设计.

发展现状和趋势有一个系统、全面的了解。对此，本文拟对已有的纵向 CDM 进行系统性地介绍并总结其发展趋势，以期国内学者更全面地了解纵向 CDM 的理论基础和应用情景，为纵向 CDA 数据分析提供理论参考。

2. 纵向认知诊断模型

经过梳理，大体可将已有的纵向 CDM 依据建模逻辑分为两大类：(1)基于潜在转换分析(latent transition analysis, LTA; Collins & Wugalter, 1992)的纵向 CDM，包括 Chen et al. (2018)、Kaya & Leita (2017)、Li et al. (2016)、 Wang, Yang et al. (2018)、Wang, Zhang et al. (2018)和 Zhang & Wang (2018)所提出的模型；(2)基于高阶潜在结构模型(higher-order latent structural model; de la Torre & Douglas, 2004)的纵向 CDM，包括 Hansen (2013)、Huang (2017)和 Zhan et al. (2019)所提出的模型。另外，鉴于 Studer (2012)和 Hansen (2013)所提出的纵向 CDM 仅适用于每个测验时间点测量 1 个属性的纵向 CDA，不符合 CDA 对多维属性进行测评的需求，本文不再介绍。同时，为避免赘述，本文也不再介绍经典的横断 CDM，比如 DINA 模型(Junker & Sijstima, 2001)、DINO 模型(Templin & Henson, 2006)和 GDINA 模型(de la Torre, 2011)等，感兴趣的读者可参阅 Rupp、Templin 和 Henson (2010)、von Davier 和 Lee (in press)及涂冬波, 蔡艳, 丁树良(2012)。

2.1. 基于 LTA 的纵向 CDM

在 CDA 中，由于潜在属性为类别变量，因此，在项目反应理论(item response theory, IRT)框架下常用的(适用于连续潜在变量)纵向建模方法(e.g., Andersen, 1985; Embretson, 1991; von Davier, Xu, &, Carstensen, 2011; Wang, Kohli, & Henn, 2015)无法直接套用在 CDM 里。对此，Li et al. (2016)将 LTA(也被称为混合隐(潜在)马尔可夫模型(mixed hidden [or latent] Markov model; Van de Pol & Langeheine, 1990))引入到 DINA 模型中，提出了适用于重复测验的 LTA-DINA 模型。类似，Kaya 和 Leita (2017)将 LTA 分别引入到了 DINA 模型和 DINO 模型中。此外，与只关注各时间点之间的变化情况(e.g., 时间点 t 到时间点 $t+1$ 的转换概率是多少)相比，也有研究者对各时间点之间的具体变化原因(e.g., 什么因素导致了时间点 t 到时间点 $t+1$ 之间的变化)更感兴趣(e.g., Wang, Yang et al., 2017)。

2.1.1. LTA-CDM

为使读者更易于理解 LTA-CDM 的建模逻辑，本文先对 LTA 进行介绍。LTA 作为潜在类别分析(latent class analysis, LCA)在重复测量(repeated measures)中的拓展，常被用于描述动态潜在变量(dynamic latent

variables; Collins & Wugalter, 1992)的在贯序阶段中的变化，比如随时间变化的态度、人格和能力。

假设测验共包含 T 个时间点，作为对 LCA 的拓展，LTA 的一般形式可表示为

$$p_{nT}=P(\mathbf{y}_{nT})=\sum_{c_T=1}^{C}P(c_T)P(\mathbf{y}_n\mid c_T)=\sum_{c_T=1}^{C}\pi_{c_T}p_{nT|c_T}\ , \tag{1}$$

式中，p_{nT} 表示到时间点 T 时 (i.e., 经过 $T-1$ 次转换)被试 n 呈现 $I\times T$ 道题目的作答结果向量 $\mathbf{y}_{nT}=(y_{n11}, y_{n21}, \ldots, y_{nIt}, \ldots, y_{n1T}, y_{n2T}, \ldots, y_{nIT})'$ 的联合概率；$\pi_{c_T}=P(c_T)$ 为到时间点 T 时的混合比例(mixing proportion)，用于描述到时间点 T 时每个类别中所包含人数占总人数的比例；C 为总类别数量(假设所有时间点上类别总数一致，因此 $c_t\le C$)；$p_{nT|c_T}=P(\mathbf{y}_{nT}\mid c_T)$ 表示到时间点 T 时归入第 c_T 类别的被试 n 呈现 $I\times T$ 道题目的作答结果向量 $\mathbf{y}_n$ 的联合概率，可进一步表示为

$$p_{nT|c_T}=P(\mathbf{y}_{nT}\mid c_T)=\prod_{t=1}^{T}\prod_{i=1}^{I}p_{nit|c_t}^{y_{nit}}(1-p_{nit|c_t})^{1-y_{nit}}\ , \tag{2}$$

式中 $p_{nit|c_t}$ 为测量模型(measurement model)，表示在时间点 t ($t\le T$)归入第 c_t 类别的被试 n 答对题目 i 的概率，作答结果 $y_{nit}\in\mathbf{y}_{nT}$。通常，LTA 假设测量模型具有参数不变性，即在任一时间点归入相同类别的被试作答相同题目的正确概率不变(实际上是假设题目参数跨时间点不变)，因此 $p_{nit|c_t}=p_{ni|c_t}$。

需要注意的是，在 LTA 中，时间点 $t+1$ 的混合比例是根据时间点 t 的混合比例以及从 t 到 $t+1$ 的转换概率(transition probabilities)计算出来的。因此，在 LTA 中仅需估计第一时间点的混合比例和不同时间点之间的转换概率即可计算出第二时间点及之后时间点上的混合比例[3]，即

$$\pi_{c_{t+1}}=\sum_{c_t=1}^{C}\pi_{c_t}\tau_{c_{t+1}|c_t}\ , \tag{3}$$

式中，$\tau_{c_{t+1}|c_t}$ 为从时间点 t 到 $t+1$，被试由第 c_t 类别转换为第 c_{t+1} 类别的概率。根据类别总数 C，从时间点 t 到 $t+1$ 有 $C\times C$ 的转换概率矩阵

[3] 这是 LTA 与 LCA 的主要区别。当数据分析者不关心被试从第 1 时间点到第 $t+1$ 时间点的依次变化情况时，则可以使用 LCA 直接估计 $\pi_{c_{t+1}}$.

$$\begin{bmatrix} \tau_{1_{t+1}|1_t} & \tau_{2_{t+1}|1_t} & \cdots & \tau_{C_{t+1}|1_t} \\ \tau_{1_{t+1}|2_t} & \tau_{2_{t+1}|2_t} & \cdots & \tau_{C_{t+1}|2_t} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{1_{t+1}|C_t} & \tau_{2_{t+1}|C_t} & \cdots & \tau_{C_{t+1}|C_t} \end{bmatrix}, \tag{4}$$

其中行为时间点 t 的类别，列为时间点 $t+1$ 的类别。对于每一位被试，从时间点 t 中的某一类别能且仅能转换到时间点 $t+1$ 中的 1 个类别，因此，式(4)中每行之和均为 1 (i.e., $\sum_{c_{t+1}=1}^{C}\tau_{c_{t+1}|c_t}=1$)。进而，整个测验经过 $T-1$ 次转换，待估计的转换概率数量为$(T-1)C(C-1)$。

将式(3)依时间点依次带入式(1)，则有

$$p_{nT}=\sum_{c_T=1}^{C}\cdots\sum_{c_2=1}^{C}\sum_{c_1=1}^{C}\pi_{c_1}\tau_{c_2|c_1}\cdots\tau_{c_T|c_{T-1}}\prod_{t=1}^{T}\prod_{i=1}^{I}p_{ni|c_t}^{y_{nit}}(1-p_{ni|c_t})^{1-y_{nit}}\ . \tag{5}$$

在 LTA 基础上，我们仅需要把测量模型设定为 CDM 即可得到 LTA-DINA 和 LTA-DINO 模型。对于 LTA-DINA 模型，有

$$p_{ni|c_t}=g_i+(1-s_i-g_i)\eta_{ni|c_t}=g_i+(1-s_i-g_i)\prod_{k=1}^{K}\alpha_{nk|c_t}^{q_{ikt}}\ , \tag{6}$$

式中，s_i 为题目 i 的失误参数，g_i 为题目 i 的猜测参数；q_{ikt} 为时间点 t 上 Q_t 矩阵中元素，$q_{ikt}=1$ 表示时间点 t 上题目 i 考查属性 k，$q_{ikt}=0$ 反之；K 为属性数量。需要强调的是，由于 LTA-DINA 和 LTA-DINO 模型仅适用于重复测验，所以在不同时间点上的 Q 矩阵完全相同，进而 $q_{ikt}=q_{ik}$； $\eta_{ni|c_t}$ 为归入第 c_t 类别的被试 n 在题目 i 上的理想作答概率；$\boldsymbol{\alpha}_{n|c_t}=(\alpha_{n1|c_t},\ldots,\alpha_{nK|c_t})'$ 为归入第 c_t 类别的被试 n 的属性向量。类似，对于 LTA-DINO 模型，则有

$$p_{ni|c_t}=g_i+(1-s_i-g_i)(1-\prod_{k=1}^{K}(1-\alpha_{nk|c_t})^{q_{ikt}}). \tag{7}$$

需要强调的是，Li et al. (2016)与 Kaya 和 Leita (2017)所提出模型并不完全相同，它们之间的主要区别在于对转换概率的建模和估计方法。对于 LTA-CDM 而言，式(4)中的转换概率实际上是属性向量(类别)水平转换概率。Kaya 和 Leita (2017)直接估计了各时间点上属性向量水平转换概率，则其待估计参数数量为$(T-1)C(C-1)$，而 Li et al. (2016)是将属性向量水平转换概率拆分为属性水平转换概率，此时需假设各属性之间的转换概率相互独立，则其待估计参数数量为$(T-1)2K$。比如，某测验仅包含 2 个时间点并考查 2 个属性，则从时间点 1 到时间点 2 针对(00)→(10)这一转换，属性向量水平转换概率描

述的是(00)→(10)的概率，即$\tau_{(10)\,|\,(00)}$；而相应的属性水平转换概率描述的是 0→1 和 0→0 的联合概率，即$\tau_{1\,|\,0} \times \tau_{0\,|\,0}$。由于 $C = 2^K$，所以属性水平转换概率的待估计参数数量少于属性向量水平转换概率的待估参数数量。

2.1.2. 一阶隐马尔可夫模型

与 Kaya 和 Leita (2017)的研究类似，Chen et al. (2018)也提出了关注属性向量水平转换概率的一阶隐马尔可夫模型(first-order hidden Markov model, FOHM)。首先，Chen et al. (2018)区分了两种属性向量水平的学习轨迹：无约束的(unrestricted)和不减少的(nondecreasing)。前者允许被试遗忘已掌握的属性，而后者假设被试一旦掌握某属性就不会出现遗忘。显然，不减少学习轨迹所需估计的属性向量水平转换概率数量远小于无约束学习轨迹的。然后，为进一步减少参数估计数量，Chen et al. (2018)又假设属性向量水平转换概率具有跨时间点不变性。因此，FOHM 本质上是对 Kaya 和 Leita (2017)所提出模型的简化。实际上，Chen et al. (2018)在 FOHM 的基础上还提出了一个高阶 FOHM 模型，即对 FOHM 引入一个一般(高阶)学习能力。但本质上讲，高阶 FOHM 仅是下文将要介绍的引入协变量的高阶隐马尔可夫模型(higher-order, hidden Markov model, HO-HMM; Wang, Yang et al., 2018)的一个特例。

2.1.3. 引入协变量的高阶隐马尔可夫模型

与上述 3 个纵向 CDM 不同，引入协变量的高阶隐马尔可夫模型(HO-HMM)更关注的是什么原因导致了被试对属性的掌握状态从时间点 t 到时间点 $t + 1$ 之间的变化，因此，不是简单的估计转换概率而是对其进行建模。与 Li et al. (2016)类似，该模型也是从属性水平转换概率入手，则对于二分属性而言，就存在 4 种转换概率$\tau_{0\,|\,0}$、$\tau_{1\,|\,0}$、$\tau_{0\,|\,1}$和$\tau_{1\,|\,1}$，分别表示 0→0、0→1、1→0 和 1→1 的概率，且前两者的和、后两者的和均为 1。为简化模型，Wang, Yang et al. (2018)假设一旦被试在时间点 t 掌握了属性，则在后续的时间点中就不会遗忘。因此，$\tau_{1\,|\,1} = 1$ 且$\tau_{0\,|\,1} = 0$，进而只需要对$\tau_{1\,|\,0}$进行建模即可，有

$$\tau_{1|0} = P(\alpha_{n(t+1)} = 1 \,|\, \alpha_{nt} = 0, \mathbf{Z}_{nt}) = \frac{\exp(\lambda_{0k} + \sum_{m=1}^{M} \lambda_{mk} Z_{nmt})}{1 + \exp(\lambda_{0k} + \sum_{m=1}^{M} \lambda_{mk} Z_{nmt})}, \tag{8}$$

该式表示被试从时间点 t 的未掌握到时间点 $t + 1$ 的掌握的转换概率由一系列协变量(covariate)导致，$\mathbf{Z}_{nt} = (Z_{nt1}, ..., Z_{ntM})'$为被试 n 在时间 t 的协变量向量，包括诸如一般学习能力、性别、社会经济地位、教育干预次数、已掌握的属性数量等，M 为协变量总数；λ_{0k} 为转换概率的截距，用于描述当所有协变量均为 0 时，转换概率的最低值；λ_{mk} 为每一个协变量对转换概率的贡献。

针对几个常见的协变量，式(8)可进一步表示为

$$\mathrm{logit}(\tau_{1|0}) = \lambda_{0k} + \lambda_{1k}\theta_n + \lambda_{2k}\sum_{l \neq k}\alpha_{nl}(t) + \lambda_{3k}\sum_{h=1}^{t}\sum_{j=1}^{I}q_{ikh}e_{ik}\ , \tag{9}$$

式中，$\mathrm{logit}(x) = \log(x / (1 - x))$；$\theta_n$ 为被试 n 的一般学习能力，该模型假设一般学习能力对所有属性在所有时间点之间的转换概率均有影响，且具有跨时间点不变性。进而学生的一般学习能力越高，则其从未掌握转换到掌握的概率越大；$\alpha_{nl}(t)$ 为在时间点 t 被试 n 已掌握的(除属性 k 外)属性数量；$\sum_{h=1}^{t}\sum_{i=1}^{I}q_{ikh}e_{ik}$ 为截止到时间点 t 被试 n 针对属性 k 已“练习”过的次数，其中 q_{ikh} 是时间点 t 之前的第 h 时间点上 Q_h 矩阵中的元素，e_{ik} 是预先设定好的“练习”收益(通常可设定为 1)。

最后，需要强调的是，FOHM 以及 HO-HMM 中假设学生不会遗忘已掌握属性或许仅适用于时间点间隔较短的情况，而当时间点间隔较长时(e.g., 一周或一个月)，被试就可能遗忘掉已掌握的属性(see, e.g., Zhan et al., 2019)。因此，准备使用这两个模型时，需要实践应用者验证该假设是否成立。

2.1.4. 引入题目作答时间的高阶隐马尔可夫 CDM

随着计算机成本的降低以及网络化程度的提高，在“互联网+测评”(张华华, 汪文义, 2016)背景下，对题目作答时间(item response times, RT)等过程数据的收集已成为了一种新常态。近两年，已有一些研究尝试将 RT 引入 CDM，以期探究引入 RT 对诊断学生学习所带来的影响(e.g., Wang, Zhang et al., 2018; Zhan, Jiao, & Liao, 2018)。

针对纵向 CDA，Wang, Zhang et al. (2018)提出了引入题目作答时间的高阶隐马尔可夫 CDM。该模型的基本建模逻辑与联合认知诊断建模框架(joint cognitive diagnosis modeling framework; Zhan et al., 2018)类似，即先对作答结果和 RT 分别建模，然后再将两个测量模型进行结合。其中针对作答结果的测量模型就是式(6)，而针对 RT 的测量模型则为

$$\log T_{nit} \sim N(\xi_i - \upsilon_n - \varphi G_{ni}(\boldsymbol{\alpha}_{n|c_t}),\ \omega_i^{-2})\ , \tag{10}$$

式中，$\log T_{nit}$ 为在时间点 t 被试 n 作答题目 i 耗时的对数，用于把正偏态分布矫正为正态分布；υ_n 为被试 n 的潜在速度；ξ_i 为题目 i 的时间强度参数，用于描述该题目对作答时间的基本要求；ω_i 为题目 i 的区分度参数，用于描述 $\log t_{ni}$ 分布的峰度；$G_{ni}(\cdot)$是被试 n 的属性轨迹(attribute trajectory)，反映了被试 n 的潜在速度变化情况，是一个组内效应(within-group effect)；φ用于量化潜在速度的变化，当$\varphi = 0$ 时，式(10)退化为对数正态 RT 模型(van der Linden, 2006)。

Wang, Zhang et al. (2018)给出了 $G_{ni}(\cdot)$的两种表达式，分别为

(1) $G_{ni}(\cdot)$被设定为示性函数

$$G_{ni}(\boldsymbol{\alpha}_{n|c_t}) = \begin{cases} 1 & \text{if } \boldsymbol{\alpha}'_{n|c_t}\mathbf{q}_i \geq \mathbf{q}'_i\mathbf{q}_i \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

该式表示当被试 n 在时间点 t 已掌握题目 i 所考查的所有属性，则其潜在速度将增加φ，变为$\upsilon_n + \varphi$。因此，学生的作答速度将会依据是否掌握题目所需的属性而被划分为两组。

(2) $G_{ni}(\cdot)$被设定为被试对属性的累加“练习”效应

$$G_{ni}(\boldsymbol{\alpha}_{n|c_t}) = \log\left[\sum_{m<t}\sum_{h}\eta^{*}_{nhm} + \sum_{q<i}\eta^{*}_{nqt} + 1\right], \tag{12}$$

式中 $\eta^{*}_{nhm} = \eta_h(\boldsymbol{\alpha}_{n|c_t}) \cdot I(\mathbf{q}'_h\mathbf{q}_i > 0)$，其中 $\eta_h(\boldsymbol{\alpha}_{n|c_t}) = \prod_{k=1}^{K}\alpha_{nk|c_m}^{q_{hkm}}$ 为在时间点 m 被归入第 c_m 类别的被试 n 对题目 h 的理想作答。则 η^{*}_{nhm} 表示当题目 h 包含第 i 题所考察的属性向量的子集时，在时间点 m 被试 n 对题目 h 的理想作答。进而，$\sum_{m<t}\sum_{h}\eta^{*}_{nhm}$ 表示在时间点 t 之前的所有时间点上被试 n 作答所有题目的累加“练习”效应。类似地 $\sum_{q<i}\eta^{*}_{nqt}$ 表示在时间点 t 上被试 n 作答题目 i 之前的题目的累加“练习”效应。另外，之所以对累加“练习”效应求对数，是为了与 log RT 在一个量尺上。

与 HO-HMM 一致，该模型在各时间点之间的属性水平转换概率为式(9)。另外，需要说明的是，不同于联合认知诊断建模框架(Zhan et al., 2018)中通过在一般潜在能力与潜在速度之间建立二元正态分布来联合 CDM 和 RT 模型，该模型是通过 $\varphi G_{ni}(\boldsymbol{\alpha}_{n|c_t})$ 把属性直接引入 RT 模型来进行联合建模的。因此，当$\varphi = 0$ 时，两个测量模型相互独立。

后续，Zhang 和 Wang (2018)又将该模型拓广到混合建模(mixture modeling)中，用于区分学生在不同时间点上的不同作答策略(i.e., 快速猜测策略还是正常解题策略)。当学生被估计为正常解题策略组，则该模型等价于 Wang, Zhang et al. (2018)的模型；而当学生被估计为快速猜测策略组，则其该模型假设被试对所有题目的正确作答概率均为 g^*，log RT 满足正态分布 $N(\mu_1, \sigma_1^2)$，且从时间点 t 到时间点 $t+1$ 属性向量保持不变。其中 g^*、μ_1，σ_1^2 是待估计参数。

2.2. 基于高阶潜在结构模型的纵向 CDM

本质上看，上面提到的这些基于转换概率的方法均是从潜在类别建模(latent class modeling)视角来分析纵向数据，且均可以被视为混合隐马尔可夫模型(mixture hidden Markov model; Vermunt, Tran, & Magidson, 2008)的特例或应用。与之不同，针对无法直接在属性水平上使用适用于连续潜在变量的纵向建模方法这一问题，Zhan et al. (2019)和 Huang (2017)利用了高阶潜在结构模型中一般潜在能力这一潜在变量的连续特性分别提出了基于多元正态分布策略(e.g., von Davier et al., 2011)和基于潜在增长模型策略(e.g., Wang et al., 2016)的纵向 CDM。

2.2.1. 基于多元正态分布的纵向高阶 CDM

在实践中，教育测评通常为高风险测验，因此，并不适合采用重复测验设计。针对更常见的锚题设计(anchor-item design)，Zhan et al. (2019)提出了纵向高阶 DINA (longitudinal higher-order DINA model, Long-DINA)模型。Long-DINA 模型共包含 3 + 1 层，其中，第一层为测量模型，用于描述各时间点上属性与题目作答之间的关系；第二层为高阶潜在结构模型，用于描述各时间点上一般潜在能力与属性之间的关系；第三层为纵向发展层，用于描述不同时间点上一般潜在能力的变化情况；另外，还包含一层特殊维度层，用于描述被试作答不同时间点上锚题之间的局部题目依赖性(local item dependence) (see, e.g., Paek, Park, Cai, & Chi, 2014)，这种局部题目依赖性可能是由记忆导致的。已有大量研究表明忽略可能存在的局部题目依赖性会影响参数估计的精度(e.g., Bradlow, Wainer, & Wang, 1999; Tao & Chao, 2016; 詹沛达, 李晓敏, 王文中, 边玉芳, 王立君, 2015)。

Long-DINA 模型[4]可被描述为：

(1) 第一层模型

$$\text{logit}(P(y_{nit}=1\,|\,\boldsymbol{\alpha}_{n|c_t},\gamma_{nm},\lambda_{i0t},\lambda_{i1t}))=\lambda_{i0t}+\lambda_{i1t}\prod_{k=1}^{K}\alpha_{nk|c_t}^{q_{ikt}}+r_{im}\gamma_{nm}, \tag{13}$$

式中，λ_{i0t}和λ_{i1t}分别为时间点 t 上题目 i 的截距和载荷；$\gamma_{nm} \sim N(0, 1)$为被试 n 的第 m 个特殊维度值，用于处理或提取锚题或重复题目之间的局部题目依赖性，各特殊维度之间相互独立。通常只有部分题目会涉及特殊维度，r_{im} 为题目 i 在第 m 个特殊维度上的区分度参数；其他参数同上。

(2) 第二层模型

$$\text{logit}(P(\alpha_{nk|c_t}=1\,|\,\theta_{nt},\delta_{0kt}))=\delta_{kt}\theta_{nt}-\beta_{kt},\ \boldsymbol{\theta}_n=(\theta_{n1},\ldots,\theta_{nT})', \tag{14}$$

式中，θ_{nt} 为时间点 t 上被试 n 的一般潜在能力；δ_{kt} 和β_{kt} 分别为时间点 t 上属性 k 的截距和载荷；所有

[4] 除 DINA 模型外，该建模逻辑亦可拓广至其他采用 logit 连接函数的 CDM 中。

一般潜在能力独立于式(13)中的特殊维度。被试的一般潜在能力和对各属性的掌握概率可随时间发生变化。此外，该方法假设潜在结构存在时间不变性，即不同时间点测验考查相同的潜在属性，则不同时间点上属性 k 的截距参数保持一致，$\delta_{kt} = \delta_k$ 且 $\beta_{kt} = \beta_k$。

(3) 第三层模型

$$\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nT})' \sim MVN_T(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{15}$$

其中，均值向量 $\boldsymbol{\mu} = (\mu_1, ..., \mu_T)'$，方差协方差矩阵

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \\ \vdots & \ddots & \\ \sigma_{1T} & \ldots & \sigma_T^2 \end{bmatrix}.$$

作为初始点和对比点，Long-DINA 模型假设第一个时间点上的一般潜在能力满足标准正态分布，因此，$\mu_1 = 0$ 且 $\sigma_1^2=1$；σ_{1T} 为时间点 1 和时间点 T 上一般潜在能力之间的协方差。

基于 Long-DINA 模型可以计算出学生总体及个体的变化情况。对于一般潜在能力，总体均值变化为 $\hat{\mu}_{t+1} - \hat{\mu}_t$，总体量尺变化为 $\hat{\sigma}_{t+1} / \hat{\sigma}_t$，个体的变化为 $\hat{\theta}_{n(t+1)} - \hat{\theta}_{nt}$。对于属性而言，每个时间点的混合比例变化情况以及个体属性向量的变化也均可以报告。另外，需要强调的是，Long-DINA 模型假设了潜在结构的跨时间点不变性，即只有不同时间点上所测评的属性是相同的时候，各时间点上的一般潜在能力才具有相同的含义。

2.2.2. 基于潜在增长模型的多水平 CDM

与 Long-DINA 模型的建模逻辑类似，Huang (2017)基于潜在增长模型提出了一种可测量属性变化的多水平 CDM。该模型的第一水平模型与 Long-DINA 模型的第一层和第二层模型保持一致，不同的是该模型的第二水平模型(i.e., Long-DINA 模型中的第三层模型)上使用了潜在增长模型而非多元正态分布，因此，有

$$\theta_{nt} = \eta_{1n} + \eta_{2n}(t-1) + \varepsilon_{nt}, \tag{16}$$

式中，η_{1n} 和 η_{2n} 分别为被试 n 的截距和载荷，也被称为成长因子(growth factors)，均为随机效应(random effect)；ε_{nt} 为被试 n 在时间点 t 上的回归残差，假设满足均值为 0 的正态分布，$N(0, \sigma^2_{\varepsilon(t)})$，且与其他

ε相互独立。进一步，为了引入学生背景信息等协变量[5]，η_{1n}和η_{2n}可进一步被设定为

$$\eta_{1n} = \nu_{10} + \sum_{m=1}^{M} \nu_{1m} \kappa_{nm} + \varsigma_{1n}, \tag{17}$$

$$\eta_{2n} = \nu_{20} + \sum_{m=1}^{M} \nu_{2m} \kappa_{nm} + \varsigma_{2n}, \tag{18}$$

式中，$\mathbf{v}_1 = (\nu_{10}, \nu_{11}, \ldots, \nu_{1M})'$ 和 $\mathbf{v}_2 = (\nu_{20}, \nu_{21}, \ldots, \nu_{2M})'$ 为在第三水平上对η_{1n}和η_{2n}的进一步解释；ς_{1n} 和 ς_{2n} 为回归残差，并假设两者满足均值向量为 **0** 的二元正态分布；$\boldsymbol{\kappa}_n$ 为协变量向量。

与 HO-HMM(式(8))类似，该模型的主要优点是对纵向 CDM 引入了协变量，试图解释导致学生成长的具体原因。但需要说明的是，与 Zhan et al. (2019)认为基于高阶潜在结构模型的纵向 CDM 应满足潜在结构的跨时间点不变性(i.e., 不同时间点所测量的属性不变)不同，Huang (2017)认为潜在结构可以随时间发生改变，并通过模拟研究探究了多水平 CDM 在后续时间点增加所测属性数量时的表现。实际上，当不同时间点测量不同属性时，尽管在符号上都可以用“θ”来表示不同时间点上的一般潜在能力，但它们的含义已经发生变化。用两个不同含义参数之间的差值，无法解释学生的发展或变化，也无法用于判断补救教学的效果。比如，我们不能说因为学生在第二次施测后的四则运算能力估计值要高于他在第一次施测后的分数运算能力估计值，所以学生(的分数运算能力)发展了且两次测验之间的补救教学是有效的。

3. 总结与展望

测评观念的变化引发了测评范式的改变，人们希望通过 CDA 来了解学生在多维度、细粒度的属性上的差异，进而有针对性地实施补救教学。近些年，基于“为学习而测评”的理念，为学生提供排名或分数的旧测评观念逐渐转向为有效促进学生学习提供诊断信息的新测评观念。然而，因横断 CDA 无法对诊断反馈及补救教学的成效进行判断，在面对促进学生发展这一诉求时就显得“虎头蛇尾”、“有始无终”。对此，纵向 CDA 成为了人们关注的焦点。为更客观合理地分析纵向 CDA 数据，近两年在国际上，研究者们提出了多个不同的纵向 CDM。然而，在国内对纵向 CDA 和纵向 CDM 的研究均显滞后。对此，本文对已有的纵向 CDM 进行了较为系统性的介绍，包括基于 LTA 的纵向 CDM 和基于高

[5] 与式(8)中的协变量不同，Huang (2017)模型中关注的是与时间点无关的协变量，比如性别。

阶潜在结构模型的纵向 CDM 两类。

基于现有文献，我们认为有关纵向 CDA 或纵向 CDM 仍有一些需要进一步深入探究的地方，值得国内相关学者们的关注，比如：

(1) 目前尚缺乏对不同纵向 CDM 的对比研究，为实践应用者选用合适的模型带来一定困难。本文将现有的纵向 CDM 划分为了两类，从模型建构视角看，基于高阶潜在结构模型的纵向 CDM 比基于 LTA 的纵向 CDM 更易于理解。但两类模型的实际应用效果或心理计量学性能还有待做出进一步对比；

(2) 近些年，随着计算机化测评的普及和虚拟测评(virtual assessment; Agard & von Davier, 2018)的出现，对可反映学生解题历程的过程数据(process data)的分析方法研究逐渐成为了心理计量学的前沿议题。在纵向 CDM 中，目前仅有 Wang, Zhang et al. (2018)利用了题目作答时间这一过程数据。如何将题目作答时间或其他类型过程数据引入纵向 CDM 中也是非常值得探讨的议题；

(3) 目前已有的纵向 CDM 均只关注到二分属性(binary attributes)。从理论上讲，多分属性(polytomous attributes; see, e.g., Chen & de la Torre, 2013)或概率态属性(probabilistic attributes; see, e.g., Zhan, Wang, Jiao, & Bian, 2018)比二分属性更为精细，更适于描述学生的成长或变化情况。如何将现有的纵向 CDM 拓广至多分属性情景也值得后续研究关注；

(4) 目前已有的纵向 CDM 均未探讨如何处理属性层级(attribute hierarchy; Leighton, Gierl, & Hunka, 2004)。从理论上讲，基于 LTA 的纵向 CDM 能够较为容易地处理该问题，即仅需在转换概率矩阵中删除不满足属性层级的属性向量即可；而基于高阶潜在结构模型的纵向 CDM 却较难以处理该问题(see, e.g., Zhan, Ma, Jiao, & Ding, 2019)，因此，如何在这类纵向 CDM 中处理属性层级是值得做进一步探讨的。

参考文献

张华华, 汪文义. (2016). "互联网+"测评：自适应学习之路. *江西师范大学学报(自然科学版), 40*, 441–455.

涂冬波, 蔡艳, 丁树良. (2012). *认知诊断理论、方法与应用.* 北京：北京师范大学出版社

詹沛达, 陈平, 边玉芳. (2016). 使用验证性补偿多维 IRT 模型进行认知诊断评估. *心理学报, 48*, 1347–1356.

詹沛达, 李晓敏, 王文中, 边玉芳, 王立君. (2015). 多维题组效应认知诊断模型. *心理学报, 47*, 689–701.

张颖, 边玉芳. (2017). 探索认知诊断研究的新思路——追踪研究中的诊断分析. *考试研究, 64*, 72–77.

Agard ,C., & von Davier, A. (2018). *The virtual world and reality of testing: Building virtual assessments*. In Jiao, H., & Lissitz, R. (Eds.), Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective (pp. 1-30). Charlotte, NC: Information Age Publishing.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). *Educational psychology: A cognitive view* (Vol. 6). New York: Holt, Rinehart and Winston.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419–437.

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2017). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*, 5–23.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131–157.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

Embretson, S. E. (1991). *Implications of a multidimensional latent trait model for measuring change*. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 184–197). Washington, DC: American Psychological Association.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis.* Unpublished doctoral dissertation, University of California, Los Angeles, CA.

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement, 77*, 369–388.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*, 181–204.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Studer, C. (2012). *Incorporating learning over time into the cognitive assessment framework.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education, 29*, 108–121.

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology, 20*, 213–247

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181–204.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). *Latent class models in longitudinal research*. In S. Menard (Ed.), Handbook of longitudinal research: Design, measurement, and analysis (pp. 373–385). Burlington, MA: Elsevier.

von Davier, M., & Lee, Y.-S. (in press). *Handbook of psychometric models for cognitive diagnosis*. New York: Springer.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary*

Journal, 23, 455–465.
Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics, 43*, 57–87.
Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives, 16*, 45–58.
Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3–14.
Wu, H.-M. (2018). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educational Psychology*. Advanced Online Publication, URL https://doi.org/10.1080/01443410.2018.1494819
Yang, X., & Embretson, S. E. (2007). *Construct validity and cognitive diagnostic assessment*. In Leighton, J., & Gierl, M. (Eds), Cognitive diagnostic assessment for Education: Theory and Applications (pp. 119–145). Cambridge, UK: Cambridge University Press.
Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262–286.
Zhan, P., Jiao, H., Liao D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*. Advanced Online Publication, URL https://doi.org/10.3102/1076998619827593
Zhan, P., Ma, W., Jiao, H., & Ding, S. (2019). A sequential higher-order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Applied Psychological Measurement*. Advanced Online Publication, URL https://doi.org/10.1177/0146621619832935
Zhan, P., Wang, W.-C., Jiao, H., & Bian, Y. (2018). Probabilistic-input, noisy conjunctive models for cognitive diagnosis. *Frontiers in Psychology, 9*: 997.
Zhang, S., & Wang, S. (2018). Modeling learner heterogeneity: A mixture learning model with responses and response times. *Frontiers in Psychology*, 9:2339.

Assessment for learning oriented longitudinal cognitive diagnosis models

ZHAN Peida PAN Yanfang LI Feiming

(College of Teacher Education, Zhejiang Normal University, Jinhua, China, 321004)

Abstract

Based on the idea of “assessment for learning" and aiming at promoting students' learning, the assessment pattern of objectively quantifying the learning status and providing diagnostic feedback has been increasingly valued. Compared with the cross-sectional cognitive diagnostic assessment, the longitudinal cognitive diagnostic assessment is more conducive to achieving the goal of promoting students' development. In order to make domestic scholars systematically understanding of the longitudinal cognitive diagnosis model (CDM), we first divided the existing longitudinal CDM into two types according to the modeling logic: one is based on the latent transition analysis and another one is based on the higher-order latent structural model. Then, the

theoretical basis and application scenarios of each model are introduced and explained one by one. Finally, four future research topics are concluded.

Key words: cognitive diagnosis; longitudinal study; latent transition analysis; latent class analysis; longitudinal cognitive diagnosis model